DOCUMENT RESUME

ED 318 778 TM 014 879

AUTHOR Rothenberg, Lori; Hessling, Peter A.
TITLE Applying the APA/AERA/NCME "Standards": Evidence for the Validity and Reliability of Three Statewide Teaching Assessment Instruments.
PUB DATE Apr 90
NOTE 58p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Construct Validity; Educational Assessment; Elementary Secondary Education; Meta Analysis; State Departments of Education; State Programs; *State Standards; *Teacher Evaluation; Testing Programs; Test Reliability; *Test Validity
IDENTIFIERS Florida; Florida Performance Measurement System; Georgia; North Carolina; *Standards for Educational and Psychological Tests; Teacher Performance Appraisal Scale; Teacher Performance Assessment Instrument

ABSTRACT

The statewide teaching performance assessment instruments being used in Georgia, North Carolina, and Florida were examined. Forty-one reliability and validity studies regarding the instruments in use in each state were collected from state departments and universities. Georgia uses the Georgia Teacher Performance Assessment Instrument. North Carolina uses the Teacher Performance Appraisal Instrument. Florida uses the Florida Performance Measurement System. The 41 studies were critiqued using the "Standards for Educational and Psychological Tests" of the American Psychological Association (APA), American Association for Educational Research (AERA), and the National Council on Measurement in Education (NCME). The focus was on evaluating the evidence found for the primary standards described in the "Standards." Results indicate that all three states needed further evidence of construct validity, criterion-related validity, and reliability when the instruments were used in different contexts. A 47-item list of references is included. Three appendices provide summaries of the competencies and functions measured by each instrument. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED318778

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

LORI F. ROTHENBERG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Applying the APA/AERA/NCME Standards: Evidence for the Validity and Reliability of Three Statewide Teaching Assessment Instruments

Lori Rothenberg, Peter A. Hessling

School of Education

University of North Carolina at Chapel Hill

Running head: APPLYING APA STANDARDS

TM014879

Paper to be presented at the April, 1990 meeting of the American Educational Research Association, Boston, MA.

Abstract

This study examined the statewide teaching performance assessment instruments being used in three southeastern states. Forty-one reliability and validity studies regarding each state's instrument were collected from state departments and universities. These studies were critiqued using the APA/AERA/NCME Standards For Educational and Psychological Tests. The focus was on evaluating the strength of the evidence found for the primary standards described in the Standards. The results of this study indicated that all three states needed further evidence of construct validity, criterion-related validity, and reliability when the instruments were used in different contexts.

Applying the APA/AERA/NCME Standards: Evidence for the Validity and Reliability of Three Statewide Teaching Assessment Instruments

Introduction

As the demand for quality education increases, the need for valid and reliable instruments to measure teaching competence and performance has become imperative. Some states have responded to this need at the district level, while others have sought to implement state-wide teacher assessment instruments. The purpose of this paper is to compare the evidence for validity and reliability of the state-wide instruments used in Georgia, North Carolina, and Florida. It is important to take a broad look at what has been said about the validity and reliability of these instruments for several reasons. They affect large numbers of practitioners. Their expanded use tends to have a political impact upon the public. Finally, they have the potential of defining what good teaching is for veteran practitioners as well as new teachers.

The use of state-wide instruments for assessing teaching performance seems to be a predominantly southeastern phenomenon, but one which has grown in national importance with the emphasis from Washington on state initiative. The Southeast has traditionally had a distrust of Federal controls. Moreover, during the early twentieth century local control of education was gradually eroded as a combination of northern philanthropy and state "progressive" efforts at consolidation and efficiency contributed to a relinquishing of local control of education and a dependency on the state for development and implementation of educational policies (see Dabney, 1936; Gatewood, 1960). State control of teacher evaluation as well as of curriculum and testing seem to be other manifestations of these historical phenomena.

While state-wide evaluation systems may have evolved, in part, for historical reasons, the kind of state assessment instruments that will be examined in this paper grew out of process-product research conducted in the 1970s (Soar, 1982; Soar Medley, & Coker, 1983). This research focused on identifying teaching practices that were associated with student achievement. Initially, much of this research was used to develop formative evaluation instruments that focused on competency, especially for neophyte or student teachers. Later, however, some instruments were developed to assess performance of all teachers, such as in North Carolina. A return to the earlier objective has been noted in some states (e.g., Florida; Johnston and Zwan, 1988).

We shall first briefly describe the Georgia, North Carolina and Florida instruments. The validity and reliability standards we used will be described, accompanied by a discussion of how the Standards apply to the results of the validity and reliability studies that were conducted in these states. This will be followed by a summary of the evidence that each state presented for the validity and reliability of its teacher assessment instruments. Finally, we will present a critique of the evidence based upon the Standards. We realize that in taking this approach we are deliberately ignoring the caution listed in the Standards of "evaluating the acceptability of a test ... on the literal satisfaction of every primary standard ..." (APA, 1985, p. 2). However, we believe the Standards are a useful heuristic and relevant when one considers the legal uses the Standards have been put to in court cases of the validity and reliability of tests.

How the Standards Were Applied

Given the political climate in which state teaching assessment instruments are often employed--especially the demand to act quickly--systematic validity and reliability studies are rare. Indeed, it would be politically naive for a

state superintendent to refer to a teaching assessment instrument as a "test," particularly when it is being employed for career ladder decisions. Nevertheless, these instruments are tests, whether norm or criterion-referenced, and should be subjected to the same kind of rigorous evaluation as any other test.

The focus of this paper is the examination of the strength of the evidence found for the primary standards described in the APA Standards. However, in some cases we shall refer to the evidence for particularly relevant secondary or conditional standards. Although the quantity of empirical studies was clearly important, the quality of evidence was also of primary importance in this paper.

Method

This review examined the state-wide teaching performance assessment instruments that have been employed in Georgia, North Carolina, and Florida. Georgia uses the Teacher Performance Assessment Instrument (TPAI); and North Carolina, the Teacher Performance Appraisal Instrument (TPAI). Florida currently has two different instruments: the Teacher Assessment and Development System (TADS), being used by the Dade County Public Schools; and the Florida Performance Measurement System (FPMS), being used in all other school systems in Florida. This review focused only on the Florida Performance Measurement System.

The studies that were reviewed for this paper were collected in the fall of 1988 and the spring of 1989 through letters and phone calls to state departments of education and university faculty in each of the three states. In each state, one person was generally found to be able to send the appropriate evaluation studies. Due to the manner of data collection, it is important to point out that this review is based upon studies that were made available to us, 41 of which are referenced in this paper. Also, reliability and validity studies are ongoing in all three states.

Instruments Reviewed

Georgia

The Georgia Teacher Performance Assessment Instrument (see Appendix 1) is designed to certify beginning teachers. It consists of eight competencies, such as "plans instruction to achieve selected objectives," "obtains information about the needs and progress of learners," and "maintains a positive learning climate" (Teacher Assessment Unit, 1985, p. 15.). Each competency is defined by several indicator statements which are further defined by descriptor statements, all of which "are statements of observable teaching behaviors" (Capie, Howard, & Cronin, 1986, p. 2). Some descriptors are associated with "key points" which provide further clarification (Teacher Assessment Unit, 1985). Each indicator is scored as acceptable or not acceptable based on the scores of its descriptors. The indicator scores are then aggregated to form competency scores. Teachers must demonstrate acceptable performance on each competency to be certified. Acceptable performance is defined by an "aggregated score of .75 over any two consecutive assessments" (Capie, 1987, p. 9). An earlier report (Capie, 1983) indicated that teachers who scored .85 on a single assessment would also be certified. This was still true in 1987. Teachers are not given samples of "good" plans "and/or specific samples as to how TPAI descriptors can be demonstrated" (Teacher Assessment Unit, 1985, p. 27). The instrument is clearly designed to assess what skills a teacher already possesses from his or her student training, and/or has developed within the first three years of teaching. A teacher is asked to submit a 7-10 day instructional unit which is examined by a three-member assessment team composed of a building level administrator, a peer teacher, and an outside observer. At least one rater must be in the same field as the teacher being rated (Padilla, Capie, & Cronin, 1986; Teacher Assessment Unit, 1985, p. 19).

Observations are scheduled in advance and each rater independently observes for an entire class period during the 7-10 day instructional period.

North Carolina

The North Carolina Teacher Performance Appraisal Instrument (see Appendix 2) was originally designed to provide evidence for the initial certification of beginning teachers. It is composed of 28 observable teaching practices that are catagorized under five teaching functions such as, "management of instructional time" and "instructional presentation" (Coop, Stuck, White, & Wyne, 1985). Later, three more functions that "are a mixed bag of research-based practices, school law, and practical necessity" (Holdzkom, 1987, p. 42) were added. However, researchers concerned with validity and reliability issues have not addressed these last three functions.

The TPAI is an observation-based rating instrument. Three observations are conducted per year, some or all by the school principal. Other observers are teachers who have been trained as raters. One of the three observations must be announced. The observer stays for an entire class period and, after reviewing observation data which "notes specific examples of the practices as demonstrated by the teacher " (Holdzkom, 1987, p. 43), holds a conference in which the teacher who was observed is both commended and given suggestions for improvement (Holdzkom, 1987).

Ratings are made on a six-point Likert type scale. The scale values are: (6) superior, (5) well above standard, (4) above standard, (3) at standard, (2) below standard, (1) unsatisfactory (Holdzkom, 1987). For a beginning teacher to progress from "initial certification" to "continuing certification" requires "at standard" performance on all functions. Those who wish to apply for a higher "Career Status," must score "well above standard" or "superior" on the

competencies (Floyd, 1985). North Carolina has designed training programs to "prepare teachers to participate actively in the evaluation plan" (Holdzkom, 1987, p. 44). This "effective teacher training" uses videotapes and other strategies to introduce teachers to all eight functions.

Florida

The Florida Performance Measurement System (Appendix 3) is composed of five domains, four of the five domains comprising the summative instrument. Each domain is also considered to be a formative instrument in and of itself. This is due to the fact that Florida legislation emphasizes improving teacher effectiveness first and evaluating teachers second (Coalition for the Development of the Florida Performance Measurement System, 1983). Because the studies we examined dealt with the summative instrument, we focused on it, rather than on the five formative domains. In the summative instrument which we examined, there are a total of 20 indicators of effective teaching and 19 indicators of ineffective teaching distributed over the four domains.

Florida law provides for three observers. From an initial reliability study, the Coalition for the Development of the Florida Performance Measurement System concluded that "...by combining at least two of the three observations, ... teacher performance can be reliably estimated using this instrument" (Coalition for the Development of the Florida Performance Measurement System, 1983, p. 22). The observers observe and record the teacher's performance while in the classroom. Observers code observed behaviors by frequency of occurrence. More behaviors result in higher scores for the effective items, and fewer behaviors result in higher scores for the ineffective items. A total score is based on the summation of scaled items. Quartile based scoring is used.

Unlike Georgia and North Carolina, Florida conducted a norming study. Two norm groups were found, elementary (grades K-5), and post elementary (grades 6-12). This was due to the different teaching methods employed by the two groups, i. e. lecture, interaction, independent seatwork, or labwork. Rather than creating separate norm groups, scoring adjustments for the two lower-scoring instructional methods proved to control adequately for this factor (Coalition, 1983))

The APA Standards and teaching performance assessment

Validity

According to the APA Standards (1985), the concept of validity refers to "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9). When evaluators judge teaching competence or performance, they are making inferences, whether or not they are using a particular instrument. When they are using an instrument, they should feel confident that the instrument does indeed measure what it purports to measure--presumably, teacher effectiveness.

Three traditional types of validation are considered: content, criterion, and construct. Content validation, as Guion (1974) has noted, seems to be a kind of construct validity. If we select a set of behaviors to observe from a universe we choose to call "effective teaching practices," we have, in effect, operationally defined effective teaching. The question then becomes, "have we really captured what effective teaching is all about?" This is a pertinent question, because a criticism of teaching assessment instruments is that the domain of practices that represents effective teaching has not been adequately sampled (Brandt, Duke, French, & Iwanicki, 1988). Standard 1.6 concerns the rationale for the content of the instrument and a definition of the universe from which this content was

derived. In other words, what is the source of the items on the assessment instruments?

The traditional measure of criterion-related validity in teacher performance measures is student achievement. Students' achievement is assumed to be in part due to effective teaching, as defined by the teaching practices measured on the instrument. Scores on the teaching performance instrument are, in effect, assumed to predict student achievement. Although Medley (1985) has taken this assumption to task and argues for what amounts to concurrent validity studies rather than predictive studies, this paper will not debate the issue at length. We will, however, consider the evidence for predictive validity of teacher performance measures.

If "student achievement" is being used as a criterion, its measure (e.g., the California Achievement Test) should be described (Standard 1.12) and contextual factors that might affect the measure of student achievement should be reported (Standard 1.13). Statistical analyses should be employed to determine the accuracy with which teacher effectiveness on several functions predicts student achievement (Standard 1.11). Suppose that "student time-on-task" is the criterion variable in a concurrent validity study. Perhaps time-on-task is being used as a surrogate for achievement. In such a study, raters will have to determine whether students are on-task or not. Therefore, their degree of knowledge of this criterion, and the training they receive to recognize it should be reported (Standard 1.14).

Suppose an instrument has been developed, tested, and found to be valid for ninth-grade inner-city English teachers using scores on a verbal achievement test as the criterion. Then the researcher wishes to use the same instrument and test in another ninth-grade inner-city English classroom 700 miles away. The

researcher can argue that the "test-use situation" has been "drawn from the same population of situations" (p. 16) as its preceding use (Standard 1.16) The same might not be said for 12th-grade rural science teachers whose students are being tested on their ability to solve scientific problems. Finally, if data is col'ected on teacher performance in the fall, but the criterion data is collected in the spring, there may be other reasons for student achievement (or lack of it) than the particular teacher behaviors that were observed. Time lags like this must be reported (Standard 1.18).

Regarding Standard 1.21, we maintain that differential prediction is not an issue for beginning teacher certification because candidates are assumed to be capable of passing the TPAI or other tests, given sufficient training. White, Smith, & Cunningham (1988) and Smith (1988) addressed this issue in demonstrating that student teachers could be coached to better execute practices on the North Carolina TPAI. However, using assessment instruments as the only indicator for merit pay or career ladder decisions could pose problems. Florida's norming study (Coalition, 1983) showed that, for their instrument, there were two distinct norm groups: elementary (K-5) and post-elementary (6-12). If this is true in other states that have not conducted norming studies and made necessary adjustments, a standard cut-score for both elementary and secondary teachers would result in more secondary teachers getting merit pay increases, given limited resources. There would, in fact, be "artifactual differences between regressions" (APA, 1985, p. 17) for elementary and secondary teachers' scores.

Finally, Standard 1.24 states: "if specific cut scores are recommended for decision making, ... the user's guide should caution that the rates of mis-classification will vary depending on the percentage of individuals tested who

actually belong in each category" (APA, 1985, p. 18). Assuming that raters are equally trained and are likely to make reliable ratings, the estimation of the errors they are likely to make may be based upon a normal distribution of master and non-master teachers. Suppose, however, that the population is not normal and that the bulk of teachers actually belongs in the non-master category. Given the raters' assumptions, the number of teachers that are classified as masters could be over-estimated.

Construct validity is a complicated issue in teaching performance assessment because not only is "effective teaching" a construct, but parts of effective teaching are constructs as well. For example, "classroom management" is often a construct for classrooms that are considered to be "orderly," another slippery concept. A construct validation would ask what the evidence is for inferring that scores received by teachers on a teaching performance instrument indicate that they are effective or ineffective teachers. In looking at teachers when being evaluated, there is some evidence that the instruments do not measure a single overall construct (Johnston and Zwan, 1988). In fact, some assessment instruments may actually measure how well teachers utilize a "canned" plan, or assess the thespian abilities of teachers and their students.

Construct validity is addressed in several standards. Standard 1.1, "evidence of validity should be presented for the major types of inferences for which the use of a test is recommended" (APA, 1985, p. 13), is applicable to the issue of competency versus quality of performance. For example, should the same instrument that is used to make initial certification decisions be used to make career-ladder decisions? If no evidence has been collected to justify this multiple usage, have cautions been issued to decision-makers (Standard 1.2)? When interpretations are being made on the basis of subscores, "the evidence

justifying such interpretation should be made explicit" (Standard 1.3, APA, 1985, p. 14). If a poor rating on a "Classroom management" subtest (or function) denies a teacher initial certification, how has this been justified? Standard 1.8 calls for construct validation. If we propose to measure teaching effectiveness, we need to demonstrate that we are indeed measuring this construct, and not something else, with our instrument. Likewise, we may need to distinguish a construct like "teaching effectiveness" from a construct like "teaching ability." Teachers who have not been effective on one or even several occasions may not be "bad" teachers, and the interpretation of their scores should not infer this. One final primary standard that must be considered regardless of the "type" of validity under study, is Standard 1.17. Should principals tend to rate all their teachers very high on a measure of teacher effectiveness (a ceiling effect), the range of scores will be restricted. If statistical adjustments are made for this ceiling effect, they must be reported, as must adjusted and unadjusted validity coefficients.

Reliability

In examining teacher performance assessment instruments, reliability is generally concerned with how consistent a teacher's ratings are from one observation to the next. Since all three states use multiple observers for at least a part of their teacher population, inter-rater reliability is an issue--that is, the consistency of the observations over raters. In addition to examining inter-rater reliability, some states have conducted other studies using generalizability theory to address issues such as dependability and stability. Dependability refers to the suitability of the instrument for making teacher certification decisions. This centers on the probability of misclassification. Stability refers to the generalizability of observational measures over occasion. That is, it addresses the

question, "...How well can the average behavior of each teacher be located relative to other teachers' average behavior?" (Ismail and Capie, 1987, p. 2).

The APA Standards clearly outline how the results of reliability studies must be presented, regardless of the aspect of reliability being examined. Estimates of relevant reliabilities and standard errors of measurement must be stated for each total score, subscore, or combination of scores that is reported (Standard 2.1). For assessment instruments, this means that reliabilities must be reported not only for the teacher's overall rating, but also for each function or indicator for which scores are reported. The population of interest and the sampling procedures that were used to obtain observations for estimating reliabilities should be described in as much detail as possible (Standard 2.2). The conditions under which results were obtained and the situations to which the results could be applied must be explained, in addition to the reporting of the appropriate statistics (Standard 2.3). For assessment instruments, this means that the researcher must describe the teachers who were assessed, where they taught, who the observers were, etc. when reporting reliability estimates.

The term "reliability coefficient" can have different meanings, depending on the types of evidence upon which the estimate is based. There are two standards that deal exclusively with reliability coefficients. First, reports of coefficients adjusted for restriction of range must be accompanied by their unadjusted coefficients. In addition, the standard deviations of the group actually tested and of the group for which adjusted estimates are given must be presented (Standard 2.4). Second, coefficients from internal analyses should not be substitutes for estimates of stability over time unless there is other evidence to support that interpretation in a certain context (Standard 2.6). One final primary standard that was examined in this paper is concerned with inter-rater

agreement. Standard 2.8 refers to the fact that when the judgment of raters or observers is pertinent, then the degree of agreement among raters must be reported. That is, the extent to which multiple raters agree on the performance of particular teachers on the same lesson must be reported.

Results

The results section of this paper is organized according to state. The evidence for content validity, criterion validity, construct validity, and reliability is described for each state. The primary standards are used as an outline for what types of evidence should be given. The strength of the evidence presented for validity and reliability in the 41 studies was ascertained from correlation coefficients, numbers of studies, qualifications of judges, and other measures, along with our own judgment.

Validity

Georgia. Content validity has been a major concern in Georgia. Several studies have been conducted that address Standard 1.6. Eight "experts' studies" were undertaken in the late 1970s and early 1980s that utilized literally thousands of opinions regarding the emerging TPAI. After the first study, which attempted to narrow the number of teacher competencies to the most essential, later studies focused on appropriateness of descriptors and indicators, the use of the instrument for certification, and minimum levels of performance (Johnson, Ellett, & Capie, 1981). Two of the later studies compared opinions of experts of different ethnic backgrounds. The results of these opinion studies, which indicated that most indicators and descriptors were indeed relevant, contributed to the final, most recent version of the TPAI. The experts in all studies were adequately described and did appear to be qualified experts. Another content validation study was conducted to validate the revised 1985 version of the TPAI.

Of the 919 educators whose responses were used, most agreed that the "categories and structure of the TPAI [were] sensible" (Capie, Howard, & Cronin, 1986, p. 27), the descriptors for the indicators were essential, the minimum standards for each indicator were adequate, the competencies were each essential requirements for certification, and the instrument could be generally applied to teachers in all teaching areas and grade levels (Capie, Howard, & Cronin, 1986). Based on the number of opinion studies, the number of experts polled, and the qualifications of the experts, Georgia has provided strong evidence for Standard 1.6 and the content validity of the TPAI.

Of the studies we have located, Georgia's criterion-related validation efforts address Standards 1.11, 1.12, 1.13, 1.14, 1.16, and 1.18 with some qualifications. We have found no discussion of any tendency toward non-normality of distribution, either of predictor or of criterion scores (Standard 1.11). It may be that distributions of predictor scores are indeed normal in all the studies, but, given the tendency towards higher ratings, increased means, and decreased variance in indicator ratings (Capie, 1982), this seems unlikely. Okey, et al., (1978) did not fully describe the criterion measure (Standard 1.12), but measures were carefully documented in later studies (Capie and Cronin, 1986; Padilla, et al., 1986). Georgia has tended to use home-grown criterion measures except in the early days of using standardized tests as measures of student achievement. In later content and construct validity studies, researchers seemed to be sensitive to the generalization problem noted in Standard 1.16 -- that "the specified test-use situation can be considered to have been drawn from the same population of situations on which validity generalization was conducted" (APA, 1985, p. 16). In our judgment, Georgia has implemented both predictive and concurrent validity studies, but has not justified their choices. All criterion-

related studies seem to be thought of as predictive studies, whether or not they are. For example, the use of student engagement as a criterion seems to be more of a concurrent validity issue than a predictive one. Georgia's validity studies do not address Standard 1.17, at least in the studies we have examined. Again, as in Standard 1.11, the issue may be moot, and statistical adjustments either have not had to be made for restriction of range, or, for some reason, adjustments were not made.

Georgia has provided the strongest, most consistent evidence for criterion-related validity when the criterion is either pupil perception of the learning environment or pupil engagement rates. There is evidence that each of these criteria are significantly related to teacher performance on the TPAI. The evidence for the link between performance on the TPAI and student achievement is weaker and less abundant. This is due, in part, to the different ways student achievement was measured.

A number of criterion-related validity studies have been made since 1978, in which pupil outcome variables have been the criteria for successful teaching performance. These criteria included pupil perceptions of the school learning environment, pupil engagement rates, and pupil achievement (Teacher Assessment Unit, 1985). Concerning the first criterion,

> ... three studies report moderate to robust (near .8) correlations of TPAI measures and pupil perceptions of the learning environment.... In general, highly rated teachers tend to be in classrooms where learners have positive perceptions of the learning environment on important dimensions known to be related to achievement (Capie and Ellett, 1982 p. 10).

As of 1982, three out of four studies had shown that almost all indicators were significantly correlated with class engagement rate, and composite scores on the

TPAI "reflecting factors or totals have predicted as much as forty percent of the variance in class engagement rates" (Capie and Ellett, 1982, pp. 13-14).

Student achievement is a problem criterion because, as Capie and Ellett (1982) note, it is not always easy to come up with a "best" method of measuring it (p. 14). Some studies employed standardized tests as a criterion, but "...success at relating the TPAI measures with standardized achievement test gains was minimal" (Teacher Assessment Unit, 1985, p. 5). For example, one study found scores on four of twenty competencies to be positively correlated with pupil achievement and two to be negatively correlated (Okey, Capie, Ellett, & Johnson, 1978). Other studies used more curriculum-relevant tests such as the Georgia Criterion Referenced Tests and achieved significant (r=.4; p<.05) results for the Learning Environment factor scores and seven competencies (Ellett, Capie, & Johnson, cited in Capie & Ellett, 1982). Teacher-made tests seem to have the best correlations with TPAI competencies (Capie & Ellett, 1982; Teacher Assessment Unit, 1985). One study found significant correlations between .29 and .50 at p<.10 (Okey, et al., 1978). But as Capie and Cronin (1986) stated, these tests are often characterized by "low level cognitive outcomes" (p. 4).

Capie and Ellett (1982) wrote, "when expected correlations do not materialize in a study of criterion-related validity, lack of reliability should be considered as one possible explanation..." (p. 16). This concern for reliability of the criterion variable is reflected in a study by Capie and Cronin (1986). The researchers sought to measure "a set of high cognitive level outcomes" (p. 4) in middle school science classes to differentiate the study from those which looked only at teacher-made tests, or looked only at elementary reading or mathematics classes. The criterion variable was measured by the Middle Grades Integrated Process Skill Test (MGIPT) with a reported Cronbach's Alpha of .89. Learner

ability was "equated" by using the Group Assessment of Logical Thinking (GALT), with a reported Alpha of .85. The mean of "teacher effects" (expected post-test scores based on the GALT subtracted from observed post-test scores based on the MGIPT) for each class was called the "teacher effectiveness index." Capie and Cronin (1986) found that five of six competencies were significantly correlated ($p<.05$) with the teacher effectiveness index. Eight of 23 indicators and 23 of 93 descriptors were found to have significant correlations ($p<.05$). The correlation of the TPAI as a whole with the teacher effectiveness index was .32. Capie and Cronin concluded that while total instrument score is more reliable, "intermediate levels of scoring such as the TPAI competencies are more desirable ... [because the total score] is a less valid indicator of effectiveness" (Capie and Cronin, 1986, p. 16). The researchers were clearly trying to balance the need for reliable predictor and criterion measures with the need for valid measures in order to arrive at meaningful criterion-related validity evidence.

As we maintained earlier, differential prediction (Standards 1.21 and 1.24) seems to be an issue for merit pay or career ladder decisions, as opposed to certification decisions. While Capie and Ellett (1982) reported that individual school systems were using the TPAI for all teachers (p. 21), this does not address what Capie and Sloan (1987) called "high stakes evaluation" (p. 2) for merit pay decisions. Rather, the use of the TPAI with experienced teachers seems to be for routine evaluations and to check that basic competencies are maintained. To our knowledge, investigations of differential prediction using the TPAI for merit pay decisions is still being studied in Georgia (see Capie and Sloan, 1987). Concerning Standard 1.24, we have found two studies that address the probability of misclassification in certification decisions. These studies will be described in the

reliability section of this paper. As for merit pay decisions, again, this is still under investigation.

We found no construct validity studies for Georgia's revised instrument, but several studies were made for the earlier version. "Teaching quality" was the construct being examined in two studies in the late 1970s (Ellett and Johnson, 1978; Ellett, Capie, & Johnson, 1979). In the first study, the TPAI, which at the time was divided into Classroom Procedures (CP) and Interpersonal Skills (IS) components, was compared to the Teacher Practices Observational Record (TPOR) and the Purdue Observational Rating Scales (ORS). The Teacher Practices Observational Record is a low-inference instrument based on both Deweyan and non-Deweyan statements reflecting teacher practices. The Purdue Observational Rating Scales is a high inference "... observation instrument used for recording teacher behavior and classroom interactions on nine separate dimensions: Warmth, Enthusiasm, Clarity, Variety, Individualization, Feedback, Cognitive Demand, Freedom, and On-Task Activity" (Ellett & Johnson, 1978, p. 2). Investigators first checked if experienced teachers recorded high scores on each instrument. Then, they computed correlations between each of the TPAI competencies and the appropriate parts of the TPOR and the ORS. Experience seemed to be associated with high scores on all three instruments. Thirty-five percent, or 7 of 20 of the TPAI-CP/TPOR results were significant ($p<.05$), with correlations ranging from .31 to .65. Sixty-four percent, or 12 of 19 of the TPAI-CP/ORS results were significant ($p<.05$), with correlations ranging from .32 to .75. The authors concluded that, as a whole, these correlations support the construct validity of TPAI-CP competencies for experienced teachers. Fifty-six percent, or five out of nine of the TPAI-IS/ORS correlations were statistically significant ($p<.05$), with correlations ranging from .37 to .44. However, none of

the TPAI-IS/TPOR correlations were significant. The authors noted that the method of scoring and administering the TPAI and the ORS are more similar than that between the TPAI and the TPOR. The authors concluded that the construct validity of the TPAI had generally been supported (Ellett and Johnson, 1978).

Ellett, et al. (1979), which was a partial replication of the 1978 Ellett and Johnson study, compared both TPAI indicators and competencies with the Purdue Observational Rating Scales. In addition to Classroom Procedures (CP) and Interpersonal Skills (IS) components, the Teaching Plans and Materials (TPM) component was also compared to the Purdue Observational Rating Scales. Of 135 TPM indicator/ORS subscale correlations, 48 or 36% were significant ($p<.05$). The correlations ranged from .30 to .60. Of 198 CP/ORS correlations, 62 or 31% were significant ($p<.05$). The correlations ranged from .30 to .61. Of 81 IS/ORS correlations, 25 or 31% were significant ($p<.05$). The correlations ranged from .30 to .61. This seemed to confirm the results of the previous study, although not to as great an extent. TPAI competency scores were also correlated with the ORS. Eighteen of 45, or 40% of the TPM competency/ORS subscale intercorrelations were significant ($p<.05$), with correlations ranging from .33 to .56. Thirty-one of 72, or 43% of the CP competency/ORS correlations were significant ($p<.05$), with correlations ranging from .30 to .59. Finally, 11 of 27, or 41% of the IS/ORS correlations were significant ($p<.05$) with correlations ranging from .31 to .58. These studies provide fairly strong evidence for the construct validity of the TPAI, to the extent that the Purdue Observational Rating Scales assesses "quality teaching" as well.

Capie, Ellett, & Johnson (1981) and Capie (1982) used factor analysis to examine the construct validity of the TPAI. Capie, et al. (1981) were interested in

the "stability of the factor structure loadings obtained from assessment data collected in different geographical areas in Georgia" (p. 1), and the stability of the factor loadings across certification levels. Both studies arrived at a three-factor solution--management, planning, and learning environment--which accounted for between 45% (1982) and 48% (1981) of the common variance. The 1981 study showed that these factors were stable for teachers from rural, urban, and metropolitan areas (Capie, et al., 1981, p. 6), and both the 1981 and 1982 studies indicated stability across certification areas.

Due to the magnitude of the between-factor correlations of the three factors (management, planning, learning environment), a strong one-factor solution was found in both studies which accounted for 38% (1981) and 36.5% (1982) of the total TPAI variance. Capie, et al. (1981) obtained data from 1,542 beginning teachers, and Capie (1982) examined ratings from 2,253 beginning teachers. The correlation between Management and Planning was .48 (both studies); between Planning and Learning Environment, .58 (1981) and .62 (1982); and between Management and Learning Environment, .63 (1981) and .59 (1982). The authors maintained that there is a strong, stable underlying structure to the TPAI which might be called "general teaching performance" (1981) or "quality teaching" (1982). Capie, et al. (1981) stated: "This analysis provides evidence for the construct validity of the TPAI. Each TPAI indicator is significantly contributing to the magnitude of a strong underlying dimension of teacher performance" (p. 8). In other words, the indicators do not seem to measure something other than what they are supposed to measure. Capie (1982) has some reservations about an instrument with 14 competencies, but only 3 factors, but for the way Georgia uses its instrument for certification, it is more important that "all indicators in a competency load on the same factor, since

'unidimensionality' should be implied for each competency" (Capie, 1982), p. 8). For the most part, indicators "within each competency do load on the same factor" (p.9), although a few do not. Capie suggests that some of the indicators may be misplaced.

From the above discussion it should be clear that Georgia has generally given strong evidence for APA Standard 1.8. One weakness, however, is that the TPAI has changed since the studies were made. Moreover, it seems rather bold to define what the instrument measures as "teaching quality" (Capie, 1982, p. 10). A better term might be "teaching as defined by obvious (low inference) behaviors."

Regarding Standards 1.1 and 1.2, Georgia has approached the issue of competent versus excellent performance gingerly. Capie and Anderson (cited in Capie and Sloan, 1987) realized after a preliminary study that the "normal" version of the TPAI may not be sufficient to distinguish the truly excellent teacher from one who is merely competent. The instrument was then changed to reflect "high levels of expertise and thus provide necessary discrimination in the instrument scores," and "the planning section of the instrument was modified considerably to make it consistent with what many experienced teachers do" (Capie & Sloan, 1987, pp. 2-3). The authors admitted that the modified instrument was a "logical" extension of the TPAI, because of sparse research on what constituted excellence in teaching as opposed to competence. Capie and Sloan (1987) examined whether the instrument successfully distinguished between competent and excellent teachers, using both qualitative and quantitative data. For the purposes of this study, 30 teachers were selected. Fifteen "excellent" teachers were selected, using multiple criteria, by the consensus of the coordinators of Georgia's Regional Assessment Centers. Fifteen

other teachers, who taught in similar contexts to the first group, were also selected. Raters were given 2 hours of training in qualitative, "holistic" data collection and, based upon their observations, came up with "assertions" regarding superior teaching (e.g., the teacher required several levels of cognitive effort from the students). Even with possible problems of initial misclassification of "excellent" teachers and the rather inadequate training observers received (which the researchers admitted), the perception of excellence by the observers, both qualitatively and quantitatively (using the revised TPAI), still seemed to be consistent and logical (Capie & Sloan, 1987).

Capie and Sloan (1987) discussed how the "assertions" related to the revised TPAI, and, indeed much of their study seems to be a justification for using this instrument to assess merit. Still, it is clear that in Georgia some attempt has been made to address the problem of the validity of multiple inferences with one instrument (competence versus incompetence; excellence versus mediocrity). However, we would like to see more studies that address the validity and reliability of this revised instrument. We do not feel that these two studies are sufficient evidence for using the instrument to assess merit.

Because Georgia makes certification decisions based upon competency scores rather than whole instrument scores, Standard 1.3 is relevant. The previously described content validity studies have lent credence to the importance of the separate competencies in the TPAI. Criterion validity studies of the ability of the indicators and descriptors to predict achievement have not been as convincing (Capie and Sloan, 1987). Capie and Sloan (1987) argue that the competencies, as "aggregates" of the descriptors and indicators, predict student achievement with greater accuracy. As previously mentioned, Capie and Cronin (1986) found the same results. Ultimately, the justification for the

interpretation of subscores seems to be a matter of common sense: should a poor planner or a poor classroom manager receive certification or merit pay, no matter how well he or she performs in other areas?

North Carolina. Claims for the content validity of the North Carolina Teacher Performance Appraisal Instrument (TPAI) seem to derive mainly from an extensive review of the process-product research on effective teaching practices. Over 600 research studies were synthesized to yield 28 teaching practices (White, Smith, & Cunningham, 1988). However, as White, et al. (1988) noted in the introductory section of their study, "little original research has been conducted to provide evidence for the validity of inferences about teaching skills made from TPAI ratings" (p. 2). Unlike Georgia, North Carolina has not consistently sought the opinions of expert judges regarding the content validity of the actual practices and functions of the instrument. Moreover, there seems to be little or no evidence for the content validity of the last three functions. A panel of four nationally recognized experts in teacher effectiveness did meet in 1988 to review the instrument. The experts suggested that the TPAI needed to be reordered somewhat, especially the last three functions. Indeed, their interpretation of current literature on effective teaching led them to suggest other functions and practices, especially those based on models of teaching other than the process-product model, "such as the coaching, modeling, cooperative and mastery learning models" (Brandt, Duke, French, Iwanicki, 1988, p. 11). However, North Carolina has at least provided some evidence to support the requisites of Standard 1.6--that is, the universe of teaching practices in the first five functions may be the universe represented by the 600 studies.

We located two criterion-related studies. In a criterion-related study conducted by White, et al. (1988), 14 student teachers taught eight 50-minute

science lessons about simple machines on 8 consecutive days. Two trained observers rated each teacher's teaching performance using the TPAI. The average of the two ratings was used as the measure of teaching performance. At the end of the eight lessons, all students took an objective achievement test (KR-20 reliability estimate = .85) covering the content of these lessons. Functions one, two, three, and five were significantly correlated with the class means on the achievement test ($p<.05$, one-tailed). The significant correlations ranged from .51 to .78. Ratings on the TPAI were related to student achievement, supporting the criterion validity of the TPAI.

Riner (1988) used 40 teacher volunteers and 400 student volunteers in a criterion-related validity study. Partial correlation coefficients were calculated between each TPAI function and student achievement measured by the California Achievement Test (CAT). The results suggested that only function eight (Non-Instructional Duties) was significantly correlated with CAT total scores at $p<.05$, with a correlation coefficient of .39. Each of the eight TPAI functions were significantly correlated with student gains on the math subtest of the CAT at $p<.05$, with coefficients ranging from .36 to .47. Also, the correlation between TPAI total score and CAT math score was .48 at the $p<.01$ level. Finally, there was no significant correlation between TPAI ratings and student achievement in reading.

White, et al. (1988) and Riner's (1988) results provided strong evidence for Standards 1.11, 1.12, 1.13, 1.14. and 1.18. It is interesting to note the differences in their chosen criteria and their results. Their findings seem to be in agreement with those of Georgia: that is, both states have found that their respective instruments have stronger relationships with content specific (or teacher made) criterion measures, as opposed to standardized tests. Standard 1.16 was not

applicable to the studies done in North Carolina. Although Standard 1.17 was not addressed, it probably should have been, given the restricted range of scores on particular functions. As White, et al. (1988) note, "... it is rare that a full range of ratings is obtained" (p. 5).

We did locate one other study that indirectly addressed criterion-related validity. The Division of Personnel Relations (1988) assessed the effect of the Career Development Program (CDP) on student achievement, using the results of the California Achievement Test. The TPAI was used as a part of the evaluation component of the CDP. Using data from 1985 to 1988, pilot units were compared to matched units selected by a third-party evaluator. It was concluded that "...students in the CDPs, taken as a group, have benefitted from more wide-spread good teaching than did students in the matched sample. While we cannot factor out how much of the achievement gains to credit with specific feature [sic] of Career Development or other innovations ... it is clear that the sixteen participating units have posted significant gains in student achievement, in both relative and absolute terms" (Division of Personnel Relations, 1988, p. 45). Specifically, "...the number of CDP units scoring below the national median [on the CAT, for grades 3, 6, and 8] declined. In the match units, the number declined only for Grade 3" (Division of Personnel Relations, 1988, p. i).

As far as we can tell, no study in North Carolina has addressed the problems of differential prediction (Standard 1.21). This lack seems to be particularly critical given the enlarged use of the TPAI which has been proposed in the Career Development Plan. Likewise, the possibility of misclassification has not been addressed (Standard 1.24). This last issue is extremely important for decisions that separate master teachers from other teachers when there are

limited funds at stake. Misclassification means, in effect, that truly effective teachers may not get the recognition and salary they deserve. Conversely, mediocre teachers may be rewarded in error. Furtwengler (1988) and the Southern Regional Education Board [SREB] (1988) did conduct evaluation activities of the Career Ladder program in North Carolina. However, their focus was not on the validity and reliability of the instrument being used. Rather, they were concerned with issues such as the impact of the School Career Development Program on "... improved teacher performance, employee satisfaction, ... community support of the programs" (Furtwengler, 1988, p. 2) and changes in schools because of these incentive programs (SREB, 1988). SREB (1988) does mention a 1988 review of the North Carolina TPAI, by a committee from outside the state, that judged the instrument suitable for certification and career levels I and II decisions. However, no empirical data were quoted.

We found few construct validity studies for North Carolina's TPAI. The TPAI has been used primarily for initial certification purposes. White, et al. (1988) and Smith (1988) have shown that the practices are "learnable" by student or beginning teachers. That is, significant improvement of ratings was observed after teachers were coached or given feedback regarding their performance. These studies lend some credence to the inference that an initially certified teacher can acquire the skills to teach, as determined by the TPAI.

Swartz, White, Stuck, & Patterson (1989) conducted an exploratory factor analysis of the TPAI, using the 28 practices in the first five functions. They arrived at a two-factor solution that accounted for 54% of the variance. They interpreted those two dimensions as being "Instructional Presentation" and "Management of Student Behavior" (Swartz, et al., 1989, p. 11). The evidence for Standard 1.8 is weak because only this one study has examined the structure of

the instrument. In fact, Swartz, et al. (1989) found that the instrument was measuring two distinct, but moderately correlated (r=.59), dimensions of teaching, as opposed to the five dimensions of teaching that are represented by the functions. Also, whether "Instructional Presentation" and "Management of Student Behavior" are sufficient indicators of "effective teaching" is not discussed.

Based on research available to us, Standard 1.1 is supported in part by the White, et al. (1988) and Smith (1988) studies. It is also supported in part by the research literature upon which the practices in the TPAI are based. However, we see no evidence that Standard 1.1 is being met when "mastery," as opposed to "competency," is being inferred. Likewise, with the proposed expanded use of the instrument, there does not seem to be evidence for Standard 1.2.

The TPAI, as Holdzkom (1987) writes, evaluates functions, not practices, due to "the synergistic effects of the practices, along with our own lack of knowledge about their relative impact" (pp. 41-42). For Standard 1.3, this hardly seems as compelling as the reasons Georgia gave for its competencies, which were based upon content and criterion validity studies.

Florida. The content validity (Standard 1.6) of the Florida Performance Measurement System (FPMS) was addressed in one study. "Content validity, which is a process of consensus of knowledgeable people, was supported through independent reviews of the literature by the research team and other knowledgeable persons" (Coalition for the Development of the Florida Performance Measurement System, 1983, p. 9). Although it was noted in the same study that further content validation was planned by having two or more nationally recognized experts in teaching effectiveness review the instrument,

we were unable to locate such a study. Even so, evidence for Standard 1.6 is acceptable.

We located four criterion-related validity studies. The Teacher Education Internship Project (1984) studied 19 teachers in mathematics and 31 teachers in social science in second, third, and fifth grade levels. Locally developed tests derived from teacher objectives were used for achievement measures. They found a significant correlation with FPMS total scores and math residualized gain scores (r=.57; p<.005, one-tailed), but a non-significant correlation between FPMS scores and social science scores (r=.22; p<.131). FPMS total scores were significantly correlated with student task engagement scores for both math (r=.42; p<.029) and social science (r=.44; p<.006).

A dissertation by P. Allen, (cited in Florida Department of Education, 1987), examined 38 eighth and eleventh grade American history teachers. Again, student achievement and student task engagement were the criteria. Student achievement was measured using a county-wide, standardized American history test developed to test the objectives defined for American history courses at the specific grade levels. Allen did not find a significant correlation between FPMS total scores and residualized gain scores (r=.13; p.220). However, a significant correlation between FPMS scores and student task engagement was found (r=.65; p<.0001).

Micceri (1986) studied 25 mathematics teachers. The measure of student achievement was the Comprehensive Assessment Program (CAP) mathematics subtest. No significant correlation between FPMS total scores and residualized gain scores was found (r=.109; p<.302). A dissertation by Crosby (cited in Florida Department of Education, 1987; Florida Department of Education, n.d.) studied 25 tenth grade biology teachers. The measure of student achievement was a test

derived from the textbook for the course, and was directly related to context and teacher objectives. Student task engagement was measured by having each observer cease teacher observation on four occasions during a lesson to note the number of students off task. This number was averaged across the four occasions and across all observations for a specific teacher. One minus the ratio of students off task was then employed as the measure. A single FPMS score was used for the reported correlations, using equally weighted "lab" and "regular" classroom observations. The equal weighting was justified since the reductions in FPMS behaviors in labs consistently paralleled the rankings of teachers between labs and regular classrooms (Florida Department of Education, 1987). A significant correlation was found between FPMS scores and student achievement (r=.430; p<.029). A significant correlation was also found between FPMS scores and student task engagement (r=.71; p<.0001).

Finally, a number of conclusions were reached based on a meta-analysis of the four aforementioned criterion-related validity studies using combined z values (Florida Department of Education, 1987). A significant relationship between FPMS scores and student achievement at the elementary level (n of studies =3; n of cases =73; z=2.44; p<.01), at the post-elementary level (n of studies =2; n of cases =62; z=1.88; p<.05), in mathematics (n of studies=2; n of cases =43; z=2.19; p<.01), and in biology (n of studies =1; n of cases =21; z=1.90; p<.05) was suggested. A significant relationship between student task engagement and student achievement at the elementary level (n of studies =2; n of cases =48; z=3.32; p<.01), and in social science/history courses (n of studies =2; n of cases -66; z=2.83; p<.01) was also indicated.

It appears that Florida has shown strong evidence for the criterion-related validity of the FPMS, when the criterion is student task engagement. The

evidence is conflicting when the criterion is student achievement. This may be due to the types of tests employed as criteria in each of the studies. In reference to the Standards, Florida has conducted studies that address Standards 1.11, 1.12, 1.13, 1.14, 1.16, and 1.18, with some small exceptions. In some cases, the tests are not completely described, so it is difficult to ascertain whether they tend toward non-normal distributions, for example. The rationale for choosing student achievement tests is clear, but it is not always made clear as to why local, district, or state achievement tests were used. The amount of time that elapsed between teacher observations and the collection of the criterion data was always clearly reported. The time lapse was also usually logical. Standard 1.17 addresses the issue of reporting adjusted and unadjusted coefficients. Florida does not make reference to this at all. Standard 1.19 asks for a rationale when choosing between a predictive and a concurrent design for criterion-related validity. Florida treats all of its criterion validity studies as though they were predictive. However, given that student task engagement is assessed at the same time as the teacher is being assessed, that portion of each study would seem to be a concurrent design.

Standard 1.21 is concerned with differential prediction studies. These would seem to apply to the validity of Master Teacher Program decisions. However, we were not able to locate studies that assess this as of yet. Standard 1.24 refers to decisions made on the basis of cut-scores. Florida has not addressed the issue of probability of misclassification yet.

We found no studies that directly fit the logic of construct validity (Standards 1.1, 1.2, and 1.8). However, Smith, Peterson, and Micceri (1987) cited the results of a factor analysis conducted using Master Teacher Program data. There were 36,000 observations of 18,000 candidates. It was reported that the results supported "the domain structure, as currently defined, and the equal

weighting of items in scoring the summative instrument" (p. 18). There is some evidence for Standard 1.3. Standard 1.3 is applicable to the FPMS since teachers are scored as being in the lowest 25%, in the middle 50%, or in the highest 25% of the normed population with regard to each item on the instrument. Then, item scores are summed to get total scores for effective and ineffective indicators. Evidence is provided for this interpretation of scores through the normative data collected in the 1982-1983 norming study. Although a proposed method of determining a cut-off for master versus non-master status with a rationale is also given, no mention is made in other studies as to whether this cut-off is indeed adopted.

Reliability

Georgia

Three reliability studies were examined regarding Georgia's TPAI. These studies addressed issues of dependability of assessment decisions, internal consistency of the TPAI, stability of the TPAI, and inter-rater agreement.

Padilla, et al., (1986) examined (a) the extent to which adding an additional observer to the assessment team influenced the reliability of assessment decisions and (b) the extent to which the type of observer (administrator, peer, external rater) affected the reliability of assessment decisions. Forty seventh-grade science teachers were studied. Generalizability theory was used for the analyses. A three facet fully-crossed design with teachers, observer types, and TPAI indicators as sources of variation was used. For each analysis, teachers were considered to be the facet of differentiation and other facets were treated as random "facets of differentiation" (Padilla, et al., 1986, p. 5). They concluded that three observers were sufficient (rho-squared = .71) and that the addition of a second external observer to make a team of four, did not add anything to the

reliability (rho-squared = .69). They also found that external observers were more consistent in the way that they scored than were other observer types, particularly administrators. The implications of this finding may be that school administrators are not very reliable raters. However, the differences were not great. Rho-squared was greater than .8 for each type. The authors note that these differences may not be as great as the day-to-day variation in the teachers' scores.

Another study (Capie, 1987) focused on dependability. The investigation, which studied 1,696 teachers who were applying for their initial certification in Georgia, was concerned with determining the suitability of the revised TPAI for making certification decisions. Generalizability theory was used to plan the analyses. A three facet, fully-crossed design with teachers, observer-types, and performance indicators as sources of variation was used. For each analysis, teachers were treated as facets of differentiation, and observer-type and performance indicators were treated as fixed facets of differentiation. The teacher by observer-type interactions were considered error and a random facet because all observers were "trained to the same criterion" (Capie, 1987, p. 4). Using Brennan's index of dependability, Capie found that using .75 as a cut-off score for each competency for certification decisions was quite acceptable. The probability of false denials on the "weakest" competency, (competency eight), is .0000000032 over the six assessment opportunities that a candidate has. In the course of the study, the generalizability coefficients of competencies five, six, seven and eight were found to be .45, .55, .59,and .56 respectively. These are less than satisfactory. Capie, however, discussed sources that could contribute to these lowered generalizability coefficients and how to approach a solution to this problem. In an earlier study, Capie (1983) found the probability of making a false denial to range from .0001 to .035 for the fourteen competencies of the earlier version of

the TPAI. This was based on two assessments in a single assessment year. Thus, both of these studies show that it is unlikely that a candidate will be misclassified.

It should be noted that Cronin and Capie (1985) found in a field test of a preliminary version of the revised TPAI that the new "essential" descriptor scoring system did not detract from the reliability of the measures. In fact, it increased the reliability (rho-squared=.65 for non-essential scoring; rho-squared=.68 for essential scoring). Since 26 teachers with a total of 104 observations were used, it was suggested that a replication study be done with more teachers under more realistic conditions. We had no indication that the new scoring system was in fact adopted.

In another study (Ismail & Capie, 1987), generalizability theory was used for the analysis of experienced teachers' TPAI assessment data. The purposes of the study were to examine (a) the effects of the number of observations on reliability, and (b) the stability of measures over time. A four facet, fully-crossed design with teachers, the 92 observation descriptors, observers, and days as sources of variation was used. Teachers were considered the facet of differentiation and all facets were treated as random. Eight pairs of teachers were studied. They found rho-squared to exceed .6 only with ten or more days of observation and twenty or more descriptors.

Cronin and Capie (1986) used generalizability theory in a study of 20 teachers observed by two observers at the same time on two different days. A D-study was conducted to determine the effects of observations made on 1, 2, 3, 4, and 5 days with one and two observers. This study employed a four facet, fully-crossed design with observer types, day of observation, and indicators as random facets of generalization and teachers as the facet of differentiation. Total

instrument scores were analyzed. Rho-squared ranged from .38 to .63 for one observer and from .53 to .76 for two observers. This provided information as to the influence of the number of days and/or observers on the decisions made. Increasing the number of days and the number sof observers clearly increased reliability.

Yap and Capie (1985) conducted a study using the preliminary version of the revised TPAI that investigated the reliability differences under two circumstances. The first situation consisted of two observers simultaneously observing a teacher on one day. The second situation consisted of the two observers observing the same teacher on different days. Twenty-three teacher volunteers were studied. Generalizability theory was used for the analysis. A three facet, fully-crossed design with teachers, observers, and performance indicators was employed. Satisfactory rho-squared values were obtained when the two observers observed on separate days (rho-squared-.599), but one day of observation did not prove to be sufficient.

Although studies of inter-rater agreement were not found for the revised instrument, Capie, Ellett, and Johnson conducted an inter-rater agreement study in 1979. Thirty teachers were observed by three or four observers. An index of exact agreement and an index of near agreement were calculated. The mean exact and near agreement rates were .98 and .86 for the Teaching Plans and Materials component, with a range of .77 to .92 for exact agreement and .85 to 1.00 for near agreement. The mean exact and near agreement rates were .87 and .98 for the Classroom Procedures component, with a range of .81 to .95 for exact agreement and .92 to 1.00 for near agreement. Finally, the mean exact and near agreement rates were .90 and .98 for the Interpersonal Skills component, with a range of .83 to .98 for exact agreement and .94 to 1.00 for near agreement. These

rates are high, but recent studies of inter-rater agreement for the revised instrument are recommended.

Georgia provides fairly strong evidence for the reliability of the TPAI. What needs to be done is an examination of inter-rater agreement for the most recently revised TPAI. In reporting the results of the reliability studies, Georgia has consistently addressed Standards 2.1, 2.2 and 2.3 satisfactorily. Standards 2.4 and 2.6 are not issues addressed. The evidence for Standard 2.8 is weak only because inter-rater agreement rates have not been reported for the revised instrument.

North Carolina

North Carolina has conducted three studies that directly address the question of reliability. Smith (1986) studied 19 mathematics, 9 science, and 13 English teachers who started teaching in an urban North Carolina school system in August of 1984. All of the teachers were middle school or secondary teachers. He found that there were no differences due to the time of day, day of the week, or site of observation for teacher ratings. However, there were significant differences related to the raters who performed evaluations and the content areas in which teachers taught. Generally, mathematics teachers scored higher than English and science teachers.

A study was conducted in 1987 regarding teacher performance ratings in pilot units of the career development program. A September, 1987 report, Performance Appraisal--Cornerstone of Career Development Plan, found from simple tallying that there were wide variations among school systems in the percentage of teachers on Career Status I and II levels (Division of Personnel Relations, 1987). This suggests that raters across systems are not consistent with respect to using the same standard: "that is, a common understanding of the

criteria and their value does not exist state-wide" (Division of Personnel Relations, 1987, p. 29). However, they also concluded that when rater errors existed, they tended to be "higher, rather than lower" ratings (Division of Personnel Relations, 1987,p. 23). They noted that the bell-shaped curve for performance evaluation ratings is seen only at the state level, not within any individual district. The researchers point out that it would be unfair to pronounce the Career Development Program as a failure on the basis of this observation. The quality of teaching performance could vary widely from system to system, meaning that the differences could be real and that teaching is valued in exactly the same way across the state. More rater training, especially for functions six, seven, and eight, and more reliability check-ups are recommended.

The last study, conducted in 1988, found that across the pilot school systems of the Career Development Program, that teachers' performance is improving (Division of Personnel Relations, 1988). The researchers state that such development was a goal of the program from the beginning. However, function five, Instructional Feedback, is still a problem. It was hypothesized that this function is perceived differently by evaluators, or that there are genuine differences among teachers within and across the districts. If the function is perceived differently by evaluators, then inter-rater reliability is suspect.

In the case of North Carolina, we were unable to locate any rigorous reliability studies other than Smith's (1986). Although we were told that some generalizability studies had been done with small samples, we were unable to locate these studies. North Carolina has weak evidence for the reliability of the TPAI. Smith's (1986) study provided evidence for standards 2.1, 2.2, 2.3, and 2.8. This was the only study that addressed the Standards. However, there are a

number of suggestions for reliability studies and some studies are in progress (Division of Personnel Relations, 1988).

Florida

In 1983, the Coalition for the Development of the Florida Performance Measurement System examined three dimensions of reliability for the summative Florida Performance Measurement System (FPMS). Inter-rater agreement, stability, and discriminant reliability were studied. Stability refers to the reliability of scores over time. Discriminant reliability refers to the degree to which the instrument effectively discriminates between teachers. Nine teachers were studied. The data were analyzed using a three-way ANOVA with observers, teachers, and situations as independent variables. The 20 effective and 19 ineffective indicators were examined separately. For the effective items, inter-rater reliability was .85, stability was .86 and discriminant reliability was .79. They also investigated the difference between reliability estimates based upon two or three observers. Three observers provided the highest estimates for inter-rater reliability (.85), stability (.86), and discriminant reliability (.79). Using two observers showed little change in the reliabilities (.82, .81, .75).

The reliability estimates for the total scale of ineffective scores were much lower. Inter-rater reliability was .47, stability was .68, and discriminant reliability was .35. The two reasons cited for the lower reliabilities was that fewer ineffective behaviors occurred in lessons and observers did not code what the teacher did not do accurately. The final recommendation was to use effective indicators for initial identification of areas requiring remediation. The ineffective indicators could be used to identify specific practices that needed to be changed.

Capie and Ellett (1987) investigated the reliability of FPMS scores with 68 teachers and a total of 136 observations. Generalizability theory was used for the analysis. The investigators employed a three facet, fully-crossed design with day of observation and teaching behaviors as facets of generalization and teachers as facets of differentiation. The rho-squared was .36. This generalizability coefficient can "...provide an index of the extent to which a score can differentiate teachers and be generalized over the set of items and days of observation" (Capie & Ellett, 1987, p. 7). The cause of the low value was explained by the large amounts of error variance associated with the teacher by day of observation effect (.031 compared to .012 for teacher effect), teacher by teaching behaviors effect (.078), and teacher by day of observation by teaching behavior effect (.402). These sources of variance suppress the generalizability of the scores. The authors do warn that these data were collected in a field study and that they do not represent actual assessment conditions. Also, the official norming population data were not available for the computation of the FPMS scores. The subjects served as their own norming population. Finally, the design was limited since "teachers were nested within observers and no indication was made of this in the data set" (Capie & Ellett, 1987, p. 4).

Florida has provided strong evidence for the reliability of the FPMS, in that three different "types" of reliability were researched and found to be quite high. Although the studies address the Standards, there are aspects of some standards that are not addressed. No errors of measurement are reported (Standard 2.1). Regarding Standard 2.2, the population of interest is implied to be teachers throughout the state. The situations to which the results could be applied is not clearly explained (Standard 2.3). Standards 2.4 and 2.6 are perhaps issues that did not need to be addressed. The evidence for Standard 2.8 was well

explained and supported in the inter-rater reliability of the effective and ineffective scales.

Conclusions

As Riner (1988) has stated, "the joint Standards for Educational and Psychological Testing of the AERA, APA, and NCME provide prudent and respected professional standards to mitigate the conflict between the public's right to protection and the teachers' rights to a fair, unbiased and valid appraisal of their work" (Riner, 1988, p. 24). There are at least two main audiences for this paper: those who provide evidence for the validity and reliability of teacher assessment instruments, and those who need to convince others that the tests are fair and truly distinguish between competent and incompetent teaching or between competent and excellent teaching. While the Standards seem to be directed primarily towards the former group, we think that the latter group should be alert to them as well. Our conclusions summarize what we feel needs to be addressed regarding the validity and reliability of three teacher assessment instruments in order to meet the AERA, APA, NCME Standards, as interpreted in the context of teacher evaluation. Perhaps, with such information, decision-makers may be better able to address the political dimensions inherent in such instruments.

Validity

All three states have demonstrated to a greater or lesser extent the content validity of their teaching assessment instruments. Georgia has devoted a great deal of energy to content validity and considers it to be essential to the validation process (Capie and Ellett, 1982). Florida, like North Carolina, has chosen to seek consensus of research literature rather than consensus of many judges.

Of the three major kinds of validity, construct validity seems to be the most problematic for the three instruments. Georgia has made the most effort to provide construct validation studies, both through comparison of instruments measuring the same construct and through factor analyses. North Carolina has implicitly looked at construct validity in studying the "teachability" of the practices on the TPAI. In addition, Swartz, et. al. (1989) have made an exploratory factor analysis of the TPAI. Confirmatory studies are under way (C. W. Swartz, personal communication, March 8, 1991). Florida seems to lack construct validity evidence entirely.

Our recommendations for construct validity for the three states under consideration are: (a) construct validity studies should be undertaken for currently used instruments; and (b) when factor analysis reveals an underlying construct in the scores of an instrument, researchers should be either more tentative in saying what that construct is, or, based on what it is not, more specific (at this writing, this recommendation only applies to Georgia).

All three states have reported criterion-related validity studies. Georgia, North Carolina, and Florida have all correlated scores from their instruments with student achievement with some success. The best measure of achievement, however, does not seem to be large standardized tests, but rather locally-produced or teacher-made tests. Georgia's example of quoting reliability coefficients for criterion measures should be followed. Given the way these instruments are scored, it would behoove all three states to be attentive to problems of restriction of range. While this issue may not mean much to teachers or the general public, the failure of an instrument to adequately distinguish mediocrity from excellence in teaching would be of concern, especially in states with career ladder and merit pay programs. North Carolina

needs to broaden its focus to include all eight competencies in criterion-related validation studies.

Because all three states seem to be on the verge of applying their instruments to the issue of mastery versus minimum competency of practicing teachers, it would benefit all three states to seriously examine problems of differential prediction, bias, and misclassification. Georgia, which conducted qualitative studies of master teachers in order to make appropriate changes in its instrument, seems to be making the most progress here.

Reliability

Reliability is still an issue for Georgia, Florida, and North Carolina. Georgia has conducted more rigorous studies, but adjusted and unadjusted reliability coefficients are not addressed. It is recommended that Georgia examine reliability coefficients for different groups of teachers, such as elementary and secondary teachers.

It is recommended that North Carolina conduct inter-rater agreement studies, studies to determine the suitability of the TPAI for making certification decisions, studies to examine the effects of the number of obsevations on reliability, and studies on the stability of measures over time. It is further suggested that North Carolina conduct reliability studies on functions six, seven, and eight of its TPAI. All three states need to consistently explain the situations and population to which their results can be applied.

Although two norm groups were found in Florida, the reliability estimates for the two groups were not reported. An examination of the suitability of the FPMS for making certification decisions should be made. Finally, Florida needs to report the standard errors of measurement around the cut scores for certification and merit pay purposes.

As time goes on, it may be especially important for all three states to continually assess reliability since the behaviors on these instruments appear to be teachable. It may be that after a period of time the variance of scores in a particular population will become quite narrow. This would be troublesome in attempting to sort out master and nonmaster teachers. This should not be a problem in assessing student teachers or beginning teachers. These teachers will generally be the lower scoring ones, until they gain mastery over the competencies. This also raises the question of how often Florida will need to renorm their instrument.

In this paper we have examined the teacher assessment instruments from Georgia, North Carolina, and Florida through the lens of the APA Standards. Although this is not the way the Standards are meant to be used, they were useful in analyzing just what evidence for validity and reliability the states have presented for their instruments, and in suggesting ways to address deficiencies.

References

American Psychological Association (1985). Standards for educational and psychological tests. Washington, D. C.: American Psychological Association.

Brandt, R. M., Duke, D., French, R., Iwanicki, E. (1988). Executive summary (Report prepared for the Education Committee of the North Carolina State Legislature), Raleigh, NC.

Capie, W. (1982, June). Structure of the teacher performance assessment instruments (TPAI) (Report No. RPB 82-1). Athens: University of Georgia.

Capie, W. (1983). Further studies of the dependability of the teacher performance assessment instruments (Report No. RPB 83-1). Athens, GA: Teacher Assessment Project, College of Education, University of Georgia.

Capie, W. (1987, April). Dependability of teacher performance measures for making certification decisions. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C..

Capie, W., Cronin, L. (1986, April). How many teacher performance criteria should there be? Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 270 465).

Capie, W., Ellett, C. D. (1982, March). Issues in the measurement of teacher competencies: Validity, reliability and practicality of Georgia's assessment program (Report No. NP 82-1). Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Capie, W., Ellett, C. (1987, April). The generalizability of two classroom observation systems for assessing merit teacher performance in Florida (Report No. CNP 87-2). Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C..

Capie, W., Ellett, C. D., Johnson, Charles E. (1981). Factor analytic investigations of beginning teacher performance data (Report No. RPB 81-10). Athens, GA: Teacher Assessment Project, College of Education, University Georgia.

Capie, W., Ellett, C. D., Johnson, C. E. (1979). Selected investigations of the reliability of the teacher performance assessment instruments (A working paper) (Report No. RPB 79-4). Athens, GA: Teacher Assessment Project, College of Education, The University of Georgia.

Capie, W., Howard, K., Cronin, L. (1986). Content validation of the revised TPAI (Report No. RPB 86-1). Athens, GA: Teacher Assessment Project, University of Georgia.

Capie, W., Sloan, P. (1987, April). An investigation using quantitative and qualitative data to distinguish "merit" teachers from a contrast group (Report No. CNP 87-5). Paper presented at the International Seminar on Teacher Education.

Coalition for the Development of the Florida Performance Measurement System (1983). Teacher evaluation project. A study of measurement and training components specified in the Management Training Act. Final report for 1982-1983. Unpublished technical report.

Coop, R. H., Stuck, G. B., White, K. P., Wyne, M. D. (1985). Carolina Teaching Performance Assessment System (CTPAS): Executive summary of the research base. Chapel Hill, NC: The School of Education, University of North Carolina at Chapel Hill.

Cronin, L. & Capie, W. (1986, April). The influence of daily variation in teacher performance on the reliability and validity of assessment data. Paper presented at the annual meeting of the American Educational Research

Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 274 704).

Cronin, L. & Capie, W. (1985, April). The influence of scoring procedures on assessment decisions and their reliability. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 265 167).

Dabney, C. W. (1936). Universal education in the South. Chapel Hill: University of North Carolina Press.

Division of Personnel Relations. (1987, September). Performance appraisal--cornerstone of a career development plan. Raleigh: North Carolina Department of Public Instruction. Unpublished technical report.

Division of Personnel Relations. (1988, November). Student achievement in Career Development Program pilot units, 1985-88. Raleigh: North Carolina Department of Public Instruction. Unpublished technical report.

Division of Personnel Relations. (1988, December). Analysis of teacher performance ratings in career development program pilot units, 1987-88. Raleigh: Department of Public Instruction. Unpublished technical report.

Ellett, C. D., Capie, W., Johnson, C. E. (1979). Validating the "Teacher Performance Assessment Instruments" with the "Purdue Observation Rating Scales": A comparable instruments investigation (Report No. 79-6). Athens, GA: Teacher Assessment Project, College of Education, University of Georgia.

Ellett, C. D. & Johnson, C. E. (1978). A study of the construct validity of the teacher performance assessment instruments: Comparable instruments investigation (A working paper) (Report No. RPB 78-13). Athens, GA: Teacher Assessment Project, College of Education, University of Georgia.

Florida Department of Education, Office of Teacher Certification. (1987). Florida Performance Measurement System predictive validity report: A meta-analysis of five predictive validity studies. Unpublished technical report.

Florida Department of Education, Office of Teacher Certification. (no date). The relationship of science teacher performance to student task engagement and achievement in classrooms and laboratory settings. Unpublished technical report.

Floyd, J. (1985, November). North Carolina comprehensive plan. Paper presented at the annual national conference of the National Council of States on Inservice Education, Denver, CO. (ERIC Document Reproduction Service No. 280 829).

Furtwengler, C. B. (1988). Evaluating career ladder/incentive programs. Atlanta: Southern Regional Education Board Career Ladder Clearinghouse.

Gatewood, W. B. (1960). Eugene Clyde Brooks: Educator and public servant. Durham: Duke University Press.

Guion, Robert M. (1977). Content validity--The source of my discontent. Applied Psychological Measurement, 1, 1-10.

Holdzkom, D. (1987). Appraising teacher performance in North Carolina. Educational Leadership, 44(7), 40-44.

Ismail, Z. & Capie, W. (1987, April). The stability of measuring teacher performance: Implications for teacher evaluation. Paper presented at the National Association for Research in Science Teaching, Washington, D. C..

Johnson, C. E., Ellett, C. D., Capie, W. (1981). Experts' opinions studies of the teacher performance assessment instruments. Report eight: Study of the 1980 edition of the TPAI (Report No. RPB 81-7). Athens, GA: Teacher Ssessment Project, College of Education, University of Georgia.

Johnston, B. J. & Zwan, J. (1988, November). Observer evaluator's use of professional judgement during the teacher evaluation process. Paper prepared for the annual meeting of the American Educational Studies Association, Toronto, Canada.

Medley, D. M. (1985, April) Issues and problems in the validation of teaching and teacher professional behaviors. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 261 085).

Micceri, T.(1986). FPMS validity studies C and D: Long term studies of elementary mathematics and post-elementary literature (unpublished technical report). Tampa, FL: University of South Florida, College of Education.

Okey, J. R., Capie, W., Ellett, C. D., Johnson, C. E. (1978). Teacher performance validation studies (A preliminary draft) (Report No. RPB 78-11). Athens, GA: Teacher Assessment Project, College of Education, University of Georgia.

Padilla, M. J., Capie, W., Cronin, L. (1986, April). The influence of team size and observer-type on the validity and reliability of assessment decisions. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Riner, P. S. (1988). A study of the criterion-related validity of North Carolina's Teacher Performance Appraisal Instrument (Doctoral dissertation, University of North Carolina at Greensboro, 1988. Not yet listed in DAI at this writing).

Smith, D. R. (1986). The general applicability of the Carolina Teaching Performance Assessment System across occasion, site, rater, and subject matter Dissertation Abstracts International, 48, 1749A.

Smith, D. R. (1988, April). The impact of clinical feedback using the North Carolina Teaching Performance Appraisal Instrument on the teaching performance of student teachers. Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

Smith, B. O., Peterson, D., Micceri, T. (1987). Evaluation and professional improvement aspects of the Florida Performance Measurement System. Educational Leadership, 44(7), 16-19.

Soar, R. S. (1982). Measures of quality in the classroom. Journal of Classroom Interaction, 18(2), 7-14.

Southern Regional Education Board [SREB] (1988, December). Is "paying for performance" changing schools? The SREB Career Ladder Clearinghouse report. Southern Regional Educational Board Career Ladder Clearinghouse, Atlanta, GA.

Swartz, C. W., White, K. P., Stuck, G. B., Patterson, T. (1989). The factor structure of ratings on the Teacher Performance Appraisal Instrument. Unpublished manuscript, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Teacher Assessment Unit (Georgia), Division of Staff Development (1985). Teacher performance assessment instruments, 1985 revision (Report No. MAN 85-1). Atlanta: Georgia Department of Education.

Teacher Education Internship Project (1984). Final report. Tampa, FL: Unversity of South Florida, College of Education.

White, K., Smith, D., Cunningham, T. (1988). Rating teaching performance: The North Carolina Teaching Performance Appraisal Instrument. Educational and Psychological Measurement, 48, 1067-1074.

Yap, K. C. & Capie, W. (1985, April). The influence of same day or separate day observations on the reliability of assessment data. Paper presented at the

annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 265 166).

APPENDIX 1 *

Georgia Teacher Performance Assessment Instrument
(This list includes only the Competencies and Indicators)

I. Plans Instruction to Achieve Selected Objectives
   1. Specifies or selects learner objectives for lessons.
   2. Specifies or selects learning activities.
   3. Specifies or selects materials and/or media.
   4. Plans activities and/or assignments which take into account learner differences.

II. Obtains Information About the Needs and Progress of Learners
   5. Specifies or selects procedures or materials for assessing learner performance on objectives.
   6. Uses systematic procedures to assess all learners.

III. Demonstrates Acceptable Written and Oral Expression and Knowledge of the Subject
   7. Uses acceptable written expression.

IV. Organizes Time, Space, Materials, and Equipment for Instruction
   8. Attends to routine tasks.
   9. Uses instructional time efficiently.
   10. Provides a physical environment that is conducive to learning.
   11. Assesses learner progress during the lesson observed.
   12. Uses acceptable written expression with learners.
   13. Uses acceptable oral expression.
   14. Demonstrates command of school subject being taught.

V. Communicates With Learners
   15. Gives explanations related to lesson content.
   16. Clarifies explanations when learners misunderstand lesson content.
   17. Uses learner responses or questions regarding lesson content
   18. Provides information to learners about their progress throughout the lesson.

VI. Demonstrates Appropriate Instructional Methods
   19. Uses instructional methods acceptably.
   20. Matches instruction to learners.
   21. Uses instructional aids and materials during the lesson observed.
   22. Implements activities in a logical sequence.

VII. Maintains a Positive Learning Climate
   23. Communicates personal enthusiasm.

24. Stimulates learner interest.
25. Demonstrates warmth and friendliness.
26. Helps learners develop positive self-concept.

VIII. Maintains Appropriate Classroom Behavior

27. Maintains learner involvement in instruction.
28. Redirects learners who are off-task.
29. Communicates clear expectations about behavior.
30. Manages disruptive behavior.

*Source: Teacher Assessment Unit, Division of Staff Development (1985), p. 15.

APPENDIX 2 •

Summary of the North Carolina Teacher Performance Appraisal Instrument
(This list only includes functions and practices)

I. Management of Instructional Time
   - 1.1 Material ready
   - 1.2 Class started quickly
   - 1.3 Gets students on task
   - 1.4 Maintains high time on task

II. Management of Student Behavior
   - 2.1 Rules - administrative and organizational
   - 2.2 Rules - verbal participation
   - 2.3 Rules - movement
   - 2.4 Frequently surveys visually
   - 2.5 Stops inappropriate behavior

III. Instructional Presentation
   - 3.1 Begins with review
   - 3.2 Introduces lesson
   - 3.3 Speaks fluently/precisely
   - 3.4 Lesson understandable
   - 3.5 Provides relevant examples
   - 3.6 High success rates - tasks
   - 3.7 High success rates - questions
   - 3.8 Brisk pace
   - 3.9 Transitions between and within
   - 3.10 Assignments clear
   - 3.11 Summarizes main points

IV. Instructional Monitoring
   - 4.1 Maintains deadlines and standards
   - 4.2 Checks during independent work
   - 4.3 Assesses performance - all
   - 4.4 Questions posed one at a time

V. Instructional Feedback
   - 5.1 Feedback - in class work
   - 5.2 Feedback - out of class work
   - 5.3 Affirms correct answer quickly
   - 5.4 Sustaining feedback

VI. Facilitating Instruction
- 6.1 Plan compatible with school-wide curricular goals
- 6.2 Uses diagnostic information to develop and revise objectives and tasks
- 6.3 Maintains accurate student records
- 6.4 Plan matches objectives, learning strategies, assessment, and student needs
- 6.5 Uses available resources to support instruction

VII. Communicating Within the Educational Environment
- 7.1 Treats all students in a fair and equitable manner
- 7.2 Interacts effectively with students, co-workers, parents, and community

VIII. Performing Non-Instructional Duties
- 8.1 Carries out non-instructional duties
- 8.2 Adheres to established laws, policies, rules and regulations
- 8.3 Follows a plan for professional development and demonstrates growth

*Sources: Coop, et al. (1985); Holdzkom, (1987).

APPENDIX 3 *

The Florida Performance Measurement System:
Summative Instrument

Instructional Organization and Development

Effective Indicators

1. Begins instruction promptly
2. Handles materials in an orderly manner
3. Orients students to classwork/maintains academic focus
4. Conducts beginning/ending review
5. Questions: academic comprehension/lesson development
6. Recognizes response/amplifies/gives corrective feedback
7. Gives specific academic praise
8. Provides for practice
9. Gives directions/assigns/checks comprehension of homework, seatwork assignment/gives feedback
10. Circulates and assists students

Ineffective Indicators

1. Delays
2. Does not organize or handle materials systematically
3. Allows Talk/activity unrelated to subject
5. Poses multiple questions asked as one, unison response
5. Poses nonacademic questions/nonacademic procedural questions
6. Ignores student or response/expresses sarcasm, disgust, harshness
7. Uses general, nonspecific praise
8. Extends discourse, changes topic with no practice
9. Gives inadequate directions/no homework/no feedback
10. Remains at desk/circulates inadequately

Presentation of Subject Matter

Effective Indicators

11. Treats concept -definition/ attributes/ examples/ nonexamples
12. Discusses cause - effect/uses linking words/applies law or principle
13. States and applies academic rule
14. Develops criteria and evidence for value judgment

Ineffective Indicators

11. Gives definition or examples only
12. Discusses either cause or effect only/uses no linking word(s)

13. Does not state or does not apply academic rule
14. States value judgment with no criteria or evidence

Communication: Verbal and Nonverbal

Effective Indicators
15. Emphasizes important points
16. Expresses enthusiasm verbally/challenges students
17.
18.
19. Uses body behavior that shows interest - smiles, gestures

Ineffective Indicators
15.
16.
17. Uses vague/scrambled discourse
18. Uses loud - grating, high pitched, monotone, inaudible talk
19. Frowns, deadpan or lethargic

Management of Student Conduct

Effective Indicators
20. Stops misconduct
21. Maintains instructional momentum

Ineffective Indicators
20. Delays desist/doesn't stop misconduct/desists punitively
21. Loses momentum - fragments nonacademic directions, overdwells

* Source: Smith, et. al. (1987).